유한패턴매칭을 이용한 자연어 질의응답 시스템†

포항공과대학교 이승우 · 이근배*

1. 서 론

질의응답(QA)은 주어진 질의에 대해, 적합한 문서가 아니라 정답 그 자체를 찾아주는 것이 목표이며, 진정한 의미에서의 정보검색(IR)에 한걸음 더 나아가려는 노력이다. 예를 들어, "타지마할은 어디에 있나?"라는 질의에 대해 "인도의 아그라"라는 정확한 응답을 제시하는 것이 질의응답 시스템이 하고자 하는 바이다.

TREC-8[14]에서 질의응답 시스템의 평가를 위한 표준 평가 문헌 집합이 마련된 이후, 많은 연구가 활발히 진행되어 오고 있다. 질의응답은 정보검색이나 정보추출 (IE), 자연언어처리 (NLP) 분야의 여러 가지 기술들을 활용할 수 있다. 우선, 정보검색의 기술들 중 단락검색 기술을 활용하여, 가능한 정답을 찾을 범위를 단락이라 일컫는 문서의 작은 조각들로 줄일 수 있다. 질의어와 정답은 대개 한 두 문장 이내에 나타나는 경향이 있기 때문이다.

일단 적합한 단락이 검색된 후에는, 그 단락들로부터 정확한 정답을 찾아내기 위해 정보추출과 자연어처리 기술들이 활용될 수 있다. 우선, 정답 유형에 대한 분류 체계를 정의하고, 그러한 정답 유형에 속하는 개체를 문서 내에서 찾아내는 기술이 필요하다. 정답 유형의 분류는 WordNet[10]과 같은 온톨로지로부터 도입될 수 있고, 정답 개체를 찾아내기 위해서는 정보추출을 위한 개체명 인식 기술이 적용될 수 있다. 물론, 기존에 정의된 개체명만으로는 다양한 정답 유형을 충족시킬 수 없으므로 추출의 대상이 되는 개체명을 가능한 모든 정답 유형으로 확장해야 한다. 여기에 유한패턴매칭 기술을 적용할 수 있다.

질의응답에 대한 대부분의 연구는 찾아진 응답의 정확도에만 초점을 맞추고 있다. 그러나 대다수의 실사용자들은 수 초간의 짧은 시간도 기다릴 만큼의 참을성을 보이지 않기 때문에 질의응답 시스템의 속도 또한 매우 중요하게 다루어야 한다. 응답 속도 향상을 위한 한 방법으로 예측 정답 색인 기술을 함께 소개한다.

본 논문의 나머지 구성은 다음과 같다. 2장에서는 질의응답에 관한 기존의 연구들을 살펴보고, 3장에서 질의응답 시스템, SiteQ의 전체적인 구조를 설명한다. 그리고 4장에서는 정답 유형을 결정하고 정답을 찾아내기 위한 유한패턴매칭을 소개한다. 5장에서는 고속의 질의응답 시스템을 위한 방법을 소개하고, TREC-10과 NTCIR-3의 질의응답 평가 문헌 집합을 사용한 평가 결과를 6장에서 보여준 후, 7장에서 결론을 맺는다.

2. 관련 연구

TREC (Text REtrieval Conference)를 통해 질의응답 평가 문헌 집합[14]이 구축된 이후, 다양한 질의응답 시스템들이 개발되어 왔다. 이 장에서는 기존의 몇 가지 연구들을 살펴보고자 한다.

질의로부터 정답 유형을 판단하는 것은 질의응답 시스템의 기본적인 구성 요소이다. 정답 유형을 판단함으로써, 정답을 찾을 검색 공간을 상당히 줄일 수 있기 때문이다. 정답 유형은 주로 WordNet과 같은 온톨로지에 정의된 개념 분류[3]나 MUC[16]에서 사용되었던 개체명 분류[4]를 기반으로 정의된다. 개체명 분류는 첫 출발은 될 수 있으나 그것만으로는 필요한 정답 유형 분류를 충족할 수 없다. 질의의 대상이 그러한 개체명으로 제한되지는 않기 때문이다. 주어진 질문의 정답 유형을 판단하기 위한 방법으로 구문 분석을 이용하는 방법[3]과 Maximum Entropy 모델을 기반으로 여러 가지 자질을 학습하는 방법[4]이 사용되었다. 구문 분석은 질의어들 사이의 관계를 구하기 위해 필요하지만, 간단한 패턴만으로도 정답 유형을 충분히 판단할 수 있다. 정답 유형은 대개 의문사로 표현되며 의문문의 구조가 비교적 정형화 되어 있기 때문이다. 학습을 위해서는 각 질문 유형에 속하는 질문 말뭉치가 충분히 있어야 하지만, 아

† 본 연구는 과학기술부 21C 프론티어사업 (지능형 로봇) 중 Human-Robot Interface 과제의 지원을 받아 연구되었음.

* 종신회원

직까지는 그러한 질문 말뭉치가 준비되어 있지 못하다.

정답을 찾을 검색 공간을 줄이기 위한 또 하나의 방법으로 단락검색 기술이 사용된다. 단락검색은 본래 정보검색 시스템의 정확도 - 특히 상위 문서의 정확도 - 를 향상시키기 위해 개발되었다. 이 기술은 단락의 정의에 따라 두 가지로 나눠 볼 수 있다. 하나는 질의에 상관없이 각 단락이 미리 고정된 정적 단락이고[15], 다른 하나는 질의에 따라 그에 맞는 단락이 정해지는 동적 단락이다[5].

마지막으로, 문서로부터 가능한 정답을 골라내는 방법으로, 여러 질의응답 시스템들은 개체명 인식기를 사용하거나[3,11] 어휘 패턴으로 기술된 규칙을 사용하였다[13]. 개체명 인식기는 인명이나 지명, 기관명 같은 고유 명칭이나, 날짜, 수치 표현 등을 찾아 내기에는 적합하지만, 질의의 대상이 그러한 개체명으로 국한되지는 않는다. [3]과 [13]에서는 질의어와 정답을 포함하는 어휘 패턴으로 상당히 정확히 정답을 찾아 낼 수 있었다. 하지만, 어휘 패턴은 적용 범위가 좁다는 단점이 있기 때문에 품사나 의미 범주와 같은 추상화를 통해 패턴의 적용 범위를 넓힐 필요가 있다. [2]에서는 문서의 깊이 있는 분석과 추론을 적용하여 정확도를 높였지만 빠른 응답 시간을 보여주지는 못했다. 반면, [13]에서는 간단한 패턴 매칭만으로도 좋은 성능을 얻을 수 있음을 보여주었다. 본 논문에서는 어휘 패턴에 비해 좀더 유연한 유한패턴매칭 기술과 빠른 응답 시간을 갖는 실용적인 질의응답 시스템, SiteQ (Web Site Question Answering System)를 소개하고자 한다.

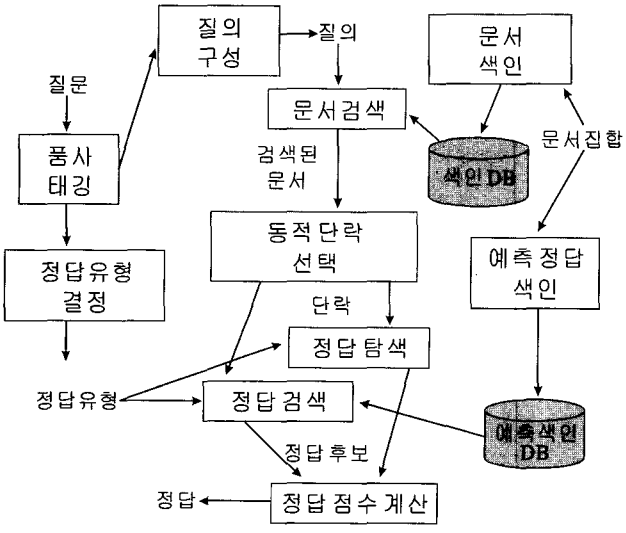


그림 1 질의응답 시스템의 전체 구조

3. 질의응답 시스템 구조

SiteQ는 특정 도메인에 국한되지 않은, 사실에 기반하여 질의에 응답하는 시스템으로 그림 1과 같은 구조를 갖는다. SiteQ는 Okapi 확률 검색 모델[12]을 적용한 문서검색 시스템, POSNIR[7]를 바탕으로 하고 있다.

보통 하나의 문서는 길고 많은 내용을 담고 있지만 정답은 질의어와 꽤 가까이 나타나기 때문에 문서 전체를 대상으로 정답을 찾는 것은 효과적이지 않다. 그래서, SiteQ에서는 우선 검색된 문서들에서 정답을 찾을 대상이 되는 단락을 골라낸다.

단락 선택이 정답을 찾을 텍스트의 공간을 좁히는 방법이라면 정답 유형 결정은 정답이 될 수 있는 후보의 범위를 개념적으로 좁히는 방법이다. 그러므로, 정확한 답을 찾기 위해서는 정답 유형 결정이 보다 더 중요하다 할 수 있다. 이를 해결하기 위해, SiteQ에서는 구문 및 의미적 처리에 바탕을 둔 유한패턴매칭을 고안하였다. 유한패턴매칭 기술은 또한 선택된 단락에서 가능한 정답을 찾을 때에도 적용된다.

문서검색 시스템을 기반으로 하는 대부분의 질의응답 시스템들은 실용적인 응답 시간을 갖지 못하고 있다. 이는 검색시에 정답을 찾기 위해 많은 텍스트를 분석하기 때문이다. 예측 정답 색인과 정답 검색은 바로 이러한 속도 문제를 극복하기 위한 방법이다. 예측 정답 색인을 적용하더라도 동적 단락 선택을 통한 정답 탐색 방법은 여전히 필요하다. 모든 질문에 대해서 정답 유형을 결정할 수 있는 것은 아니기 때문이다. 예를 들어, "What do bats eat?"나 "What is done with worn or outdated flag?"와 같은 질문의 경우, 정답 유형을 딱히 결정하기란 쉽지 않다. 이런 질문에 대해서는 검색시에 정답 탐색을 통해 정답을 구한다. 찾은 정답 후보들은 일치된 질의어의 수와 가중치, 질의어와 정답 후보 사이의 거리 등을 바탕으로 점수가 부여된다.

4. 견고한 구문-의미 처리를 위한 유한패턴매칭

이 장에서는 유한패턴매칭 기술을 소개한다. 어휘 의미 패턴 (Lexico Semantic Pattern: LSP)은 어휘 그 자체와, 품사, 구문 범주, 의미 범주로 표현될 수 있다. 어휘-의미 패턴은 각 어휘를 구문 및 의미 수준으로 일반화할 수 있기 때문에 어휘만으로 표현하는 패턴에 비해 보다 유연하다. 따라서, 필요한 패턴의 수를 줄일 수 있으며 복잡한 구문 및 의미적 언어 현상을 다룰 수 있는 표현력이 증가된다. 그림 2는 품사와 의미 범주에 의한 일반화의 예를 보여준다. 'NPPS'와 'NPPG'는 사람의 성과 이름을 가리키는 품사 기호이고, '@position'은 사람의 직업 혹은 역할을 가리키는 의미 범주이다[9].

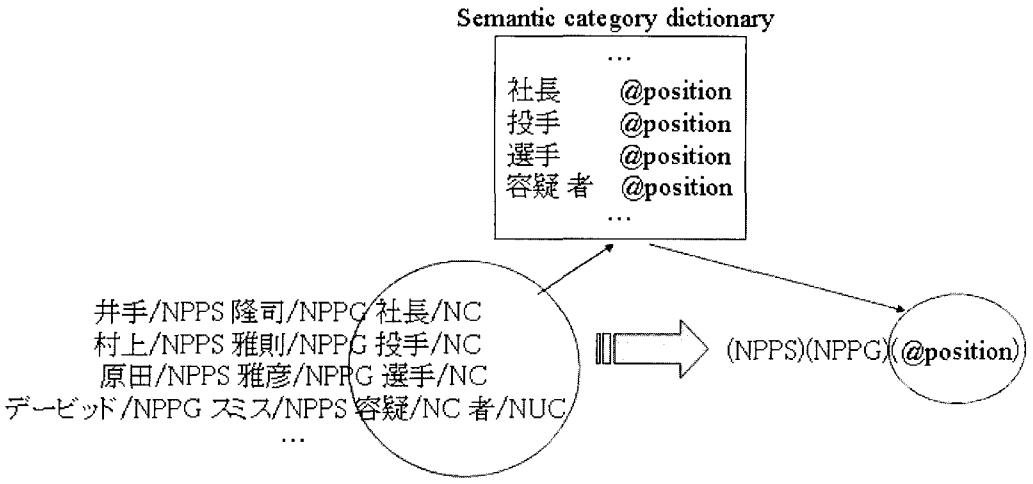


그림 2 의미 일반화에 의한 어휘-의미 패턴의 예 (일본어의 경우)

어휘를 의미 범주로 일반화하기 위해 68개의 의미 범주를 정의하고 수많은 웹 사이트와 사전으로부터 각 범주에 속하는 약 25만 개의 어휘를 수집하였다. 주요 의미 범주로는 사람, 장소, 학교, 도시, 회사, 새, 약품 등이 포함된다.

유한패턴매칭 기술은 질의응답 시스템, SiteQ의 두 부분, 즉, 정답 유형 결정과 정답 후보 탐색에서 사용된다.

4.1 정답 유형 결정

TREC QA task에서 사용된 질문들을 참조하여 62개의 정답 유형을 정의하고, 유한패턴매칭 기술을 사용하여 각 질문을 정답 유형별로 분류하는 방법을 개발하였다. 대개 의문사가 정답 유형을 결정하는 데 중요한 요소이기는 하지만 의문사 또한 중의성이 있기 때문에 그것 만으로는 부족하다. 예를 들어, "東京ディズニーランドはどこにありますか。(도쿄 디즈니랜드는 어디에 있습니까?)"에서 'どこ'는 '장소'를 가리키지만, "日本サブウェイはどこの会社の子会社ですか。(일본 서브웨이는 어느 회사의 자회사입니까?)"에서는 '회사'를 가리킨다. 즉, 의문사와 함께 주변의 문맥도 함께 고려해야 한다. 'どこにあり'와 'どこの会社'가 두 질문의 정답 유형을 말해주는 핵심구이며 어휘 의미 패턴으로 다음과 같이 표현된다.

(どこ)(に)(ある) → 1|3|location
(どこ)(の)(%company) → 1|3|company

'%company'는 '会社'의 유의어를 가리키는 의미 범주이다. 화살표의 오른쪽에서 첫 번째 숫자는 어휘-의미 패턴에서 의문사(どこ)의 위치를 가리키며, 두 번째 숫자는 어휘-의미 패턴의 구성 요소의 수를 가리킨다. 마지막 항목은 패턴이 매칭된 경우의 정답 유형(location 혹은 company)을 가리킨다. 하나 이상의 정답 유형이 가능할 경우에는 수직선(|)을 사용하여 나열될 수 있다. 정답 유형 결정을 위한 어휘-의미 패턴 규칙의 다양한 예를 표 1에 실었다.

표 1 정답 유형 결정을 위한 어휘-의미 패턴 규칙의 예[9]

의문사	질문 예	LSP 규칙
何/なん/なに (무엇/무슨)	"夏目漱石の名作は何ですか。" (소우세키 나츠메의 명작은 무엇인가?) "千葉県の縣廳所在地は何市ですか。" (치바현의 현청 소재지는 무슨 시인가?)	(%work)(は)(何)(です) → 3|4|movie|book|music (は)(何)(%city)(です) → 2|4|city
誰/だれ/どなた (누구)	"大学審議会の会長は誰ですか。" (대학 심의회의 회장은 누구인가?)	(@position)(は)(誰)(です) → 3|4|person
どこ/何処/何所, どちら/どっち (어디/어느)	"タージ・マハールはどこにありますか。" (타지 마할은 어디에 있는가?)	(どこ)(に)(ある) → 2|3|location
いつ (언제)	"米ソの冷戦が終わったのはいつですか。" (미소 냉전이 종결된 것은 언제인가?)	(は)(いつ) → 2|2|date

표 2 정답 후보 탐색을 위한 어휘-의미 패턴 규칙 예[9]

텍스트	LSP 규칙
30歳 (*30세*) 夏目房之介さん(47) (*후사코레카이 나츠메씨(47)*)	(@number)(@unit_age) → age|1|2|2|1.0 (@person)(NUP)(()(@number)()) → age|4|4|5|1.0
夏目漱石の「こころ」 (*나츠케의 「코코로」*) …主演の「L. A. コンフィデンシャル」(*...주연의 「LA 컨피덴셜」*)	(@person)(の)(@bracket) → book|3|3|3|0.85 (主演)(の)(@bracket) → movie|3|3|3|0.9
バイスフロクさん (바이수후로쿠씨) 村上雅則投手 (마사노리 무라카미 투수)	(@np)(NUP) → person|1|1|2|0.85 (NPPS)(NPPG)(@position) → person|1|2|3|1.0
スカイマークエアラインズ (本社・東京) (*스카이 마크 에어라인즈 (본사・도쿄)*)	(K)(()(本社)(・) → company|1|1|4|0.9

4.2 정답 후보 탐색

문서에서 정답의 대략적인 위치가 아니라 정확한 경계를 찾아내는 것이 중요하다. 이를 위해 어휘-의미 패턴을 사용할 수 있다. 먼저, 각 정답 유형에 해당하는 단어들을 여러 웹 사이트와 사전으로부터 충분히 수집하고 이 단어들을 사용하여 마이니치 신문 기사[1]에서 나타나는 좌우 문맥들을 수집하였다. 수집된 단어와 주변 문맥을 결합하여 정답 후보 탐색을 위한 어휘 의미 패턴을 구축하였다. 그림 3은 정답 유형 'person'에 대한 어휘-의미 패턴 규칙을 구축하는 과정을 자세히 보여준다. 예를 들어, 인명 '村上/NPPS 雅則/NPPG (*마사노리 무라카미*)'과 오른쪽 문맥 'さん/NUP (*씨*)'로부터 다음과 같은어휘-의미 패턴 규칙이 만들어진다.

(NPPS)(NPPG)(NUP) → person|1|2|3|1.0

이 규칙의 오른편은 다섯 항목으로 구성되어 있는데, 첫째는 정답 유형을 가리키며 둘째와 셋째는 각각 어휘-의미 패턴에서 정답의 시작 위치와 끝 위치를 가리킨다. 넷째는 패턴의 구성 요소 수를, 다섯째는 규칙의 신뢰도를 가리킨다. 이 신뢰도 값은 규칙을 텍스트에 적용하였을 때 올바르게 적용된 회수를 세어서 구하였다. 표 2에 정답 후보 탐색을 위한 어휘-의미 패턴 규칙의 다양한 예를 실었다. 실험에서는 약 500개의 규칙을 만들어 사용하였다.

4.3 정답 후보 점수 계산

각 정답 후보의 점수(AScore)는 단락에서 중복을 제외한 매칭된 단어의 수와 매칭된 단어들 사이의 거리를 기준으로 계산된다. 이는 정답이 매칭된 질의어와 가까이 나타날 확률이 높다는 가정을 바탕으로 한다. 점수 계산에는 다음의 네 가지 수치를 사용되며, 단락의 점수(PScore)와 결합된다.

- LSPwgt: 적용된 어휘-의미 패턴 규칙의 신뢰도
- qtuc: 중복을 제외한 질의어의 수
- ptuc: 단락에서 매칭된 중복을 제외한 질의어의 수
- avgdist: 정답 후보와 매칭된 질의어 사이의 거리의 평균 값

$$AScore = \beta \times PScore + (1-\beta) \times \frac{LSPwgt}{qtuc} \times \frac{ptuc^2}{(ptuc + avgdist)} \quad (1)$$

β는 상수 값으로 실험에서는 0.5를 사용하였다.

5. 고속의 질의응답 시스템

대부분의 질의응답 시스템에 관한 연구가 찾은 정답의 정확도에만 초점을 맞추고 있지만, 시스템의 속도 또한 중요하다. 응답 시간을 줄이는 가장 좋은 방법의 하나는 정답을 찾는데 필요한 작업들을 미리 계산해 두는 것이다. Prager[11]가 제시한 것처럼, 정의된 정답 유형에 속하는 단어들을 질문이 주어지기 전에 미리 찾아 둠으로써, 정답을 찾는데 걸리는 계산 시간을 상당히 줄일 수 있다. 그러나, 정답 후보와 질의어가 들어 있는 단락을 선택하는데 걸리는 시간은 여전히 남아 있다. 그러한 계산 시간까지 줄이기 위해서는 정답을 예측하고 그 정답을 유도하는 질문에 포함될 질의어들도 미리 예측하는 방법이 필요하다. 질의어는 대개 정답과 가까이 나타나기 때문에 예측 질의어가 나타날 범위를 예측 정답을 포함하는 작은 윈도우로 제한할 수 있고 이 윈도우는 앞서 설명한 동적 단락을 선택하는 방법으로 구할 수 있다. 그러면, 단락에서 각 단어의 가중치는 단어 빈도수와 역문헌 빈도수, 단락에서의 단어 빈도수와 문서에서의 단어 빈도수의 비율, 단어와 정답 사이의 거리를 기준으로 계산될 수 있다. 그러므로, 각 예측 정답은 각 정답 단락에서의 단어와 그 단어의 가중치로 표현될 수 있으며, 문서를 그 문서에 나타난 단어로 색인하는 것과 마찬가지로 예측 정답도 색인할 수 있다[6]. SiteQ 시스템에서 예측 정답 색인 기술을 사용함으로써 문서검색 시스템에 견줄만한 응답 속도로 정답을 찾는 것이 가능하다.

예측 정답 색인은 예측 가능한 정답 유형에 대해서는 적합하지만, 모든 질문의 답을 예측할 수 있는 것은 아

니다. 질문이 예측 가능한 정답을 요구하는 것이면 예측 정답 색인 DB에서 정답을 검색하고, 그렇지 않으면 동적 단락 선택과 정답 후보 탐색 과정을 거쳐 정답을 찾아야 한다. 그림 1의 전체 시스템 구조에서 정답 후보 탐색 모듈과 예측 정답 색인 모듈이 모두 필요한 이유가 여기에 있다.

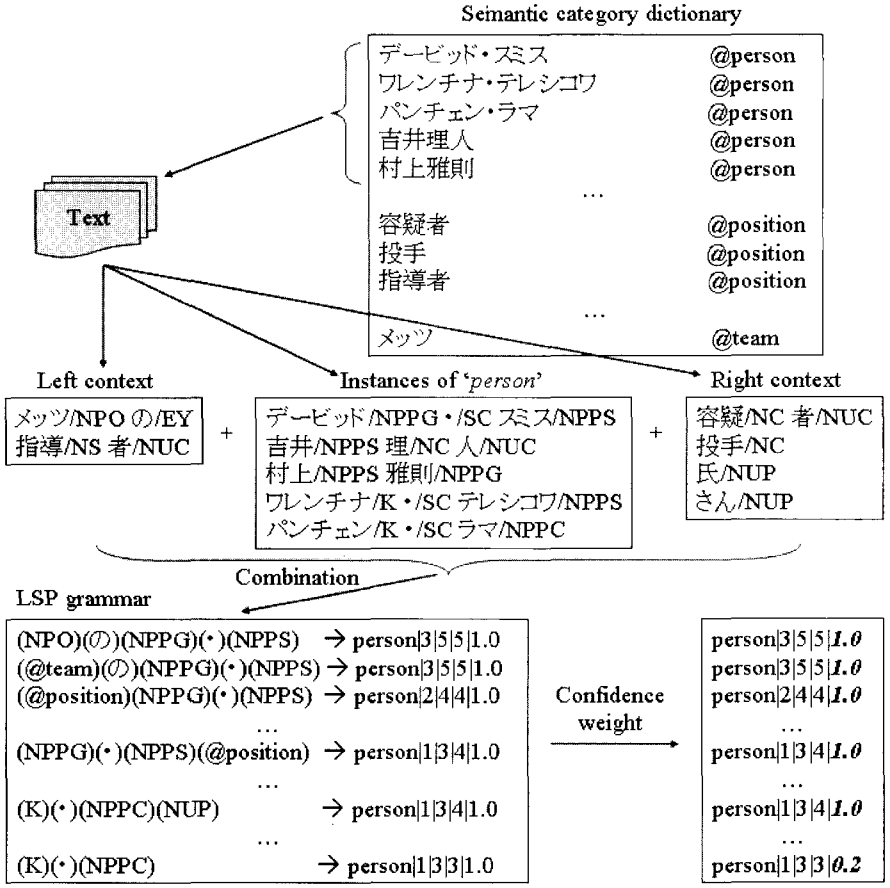


그림 3 정답 유형 'person'에 대한 어휘-의미 패턴 규칙 구축

6. 평 가

6.1 실험 환경

SiteQ 시스템의 성능을 평가하고 사용된 기술들의 효과를 살펴보기 위해 NTCIR-3 QAC 평가 문헌 집합을 사용하였다. 이 평가 문헌 집합은 23만 개의 98~99년도 마이니치 신문 기사와 National Institute of Informatics (NII)에서 제공한 200개의 질문으로 구성되어 있다. NTCIR-3 QAC task[1]는 TREC QA task와 비슷한데, 일정 길이의 정답 문자열이 아닌, 정확한 정답 그 자체를 요구하며 상위 5개의 정답을 제출할 수 있다. 각 질문에 대한 평가 점수는 제출된 5개의 정답 후보 중에서 첫 번째 정답의 순위를 역수(Reciprocal Rank) 취하여 계산된다. 만약 5개의 정답 후보 중에 정답이 없으면 점수는 0이 된다. 전체 점수는 정답이 존재하지 않는 5개의 질문을 제외한 195개 질문에 대한 점수의 평균 값 (Mean Reciprocal Rank : MRR)으로 계산된다.

6.2 NTCIR-3 QAC 데이터에 대한 실험

문서검색 시스템 (POSNIR)로부터 상위 1000개의 문서를 선택하고, 동적 단락 선택을 통해 상위 500개의 단락을 골라서 상위 5개까지의 정답 후보를 생성하였다. 이에 대한 평가 결과는 표 3에 나와 있다. 200개의 질문(question)에 대해 288개의 정답(answer)이 가능한데, SiteQ는 994개의 정답 후보를 제출(output)하여 183개의 정답(correct)을 찾았다. 재현율(Recall)과 정확도(Precision)는 이들 수치로부터 계산되며 F-measure는 (2×Recall×Precision) / (Recall + Precision)로 계산된다. SiteQ의 MRR 값은 0.608을 기록했으며 이는 참가한 15개 시스템 중에서 가장 좋은 성능이었다[1,8]. 표 4는 첫 번째 정답의 순위별로 질문의 수를 보여주고 있는데, 질문의 약 절반에 대해 첫 번째 순위에 정답을 제시하였음을 알 수 있다.

표 3 NTCIR-3 QAC 데이터에 대한 SiteQ의 성능

Question	Answer	Output	Correct
200	288	994	183
Recall	Precision	F-measure	MRR
0.635	0.184	0.285	0.608

표 4 첫 번째 정답의 순위별 질문의 수

Rank	# of questions
1	98
2	27
3	14
4	6
5	4

6.3 TREC-10 QA 데이터에 대한 실험

동일한 방법을 영어를 대상으로 한 질의응답 시스템에도 적용하였다[7]. 동적 단락 선택과 어휘-의미 패턴과 같은 기술들이 동일하게 적용되었고 NIST에서 제공한 각 질문당 상위 1000개의 검색된 문서를 사용하였다. 500개의 질문에 대한 평가 결과는 표 5에 실려 있다. 성능은 평균에 비해 훨씬 좋지만, 일본어에 대한 결과에 비해 상대적으로 낮은데, 그 이유로는 평균 성능의 차이로 볼 때 NTCIR-3 QAC의 질문이 상대적으로 TREC-10의 질문에 비해 쉬웠고, 자체적으로 색인 DB를 구성하지 않았던 것이 역문헌 빈도수(*idf*)와 같은 유용한 문서의 통계치를 사용할 수 없어서 선택된 동적 단락의 질이 떨어뜨리는 결과를 초래했기 때문인 것으로 생각된다. 이 평가 결과를 통해, SiteQ의 방법론이 특정 언어에 종속되지 않고 다른 언어에도 쉽게 확장될 수 있음을 확인할 수 있다.

표 5 TREC-10 QA 데이터에 대한 SiteQ의 성능

순위	질문수 (strict)	질문수 (lenient)	Avg. of 67 runs
1	121	124	88.58
2	45	49	28.24
3	24	29	20.46
4	15	16	12.57
5	11	14	12.46
No	276	260	329.7
MRR	**.320**	**.335**	**.234**

6.4 예측 정답 색인의 효과

예측 정답 색인 기술의 효과를 조사하기 위해, 예측 정답 색인을 사용한 경우와 그렇지 않은 경우의 응답 시간을 비교하였다. 예측 정답 색인을 빈번한 정답 유형들(사람, 장소, 기관, 날짜)에 적용하여 예측 정답 색인 DB를 미리 구성하였다. 이 정답 유형에 속하는 106개의 질문에 대해 Dual 펜티엄3 CPU에 512MB의 메모리를 갖는 리눅스 서버에서 실험을 수행했을 때, 평균적으로 각 질문당 0.56초의 응답 시간을 보였다. 예측 정답 색인을 사용하지 않은 경우의 평균 응답 시간이 8.74초인 것에 비추어 예측 정답 색인 기술을 적용함으로써 질의응답 시스템의 응답 시간을 상당히 단축시킬 수 있음을 알 수 있다.

7. 결 론

본 연구에서, 질의응답 시스템의 성능을 높이기 위한 세 가지 효과적인 기술을 소개하였다. 동적 단락 선택을 통해 검색 공간을 좁히고 어휘-의미 패턴을 사용하는 견고한 유한패턴매칭 기술을 통해 질문의 정답 유형을 결정하고 텍스트로부터 정답 후보를 탐색하였다. 실험을 통해 선택된 단락들 중 상위 10~50개 만을 사용하는 것으로 정답 후보 탐색을 위해 충분하다는 것을 알았다. 예측 정답 색인은 일종의 캐시 역할을 수행함으로써 질의응답 시스템의 빠른 응답 시간을 보장할 수 있었다.

질의응답은 진정한 의미의 정보검색으로 한 걸음 다가선 기술임은 분명하다. 그러나 현재까지 정확도나 속도 측면에서 모두 만족스러운 결과를 보이고 있지는 못하다. 무엇 보다 중요한 것은 질문을 던지는 실사용자들 스스로가 자신의 정보욕구를 적절히 표현하는데 어려움을 겪고 있다는 점이다. 다시 말해, 질문을 표현할 때, 완전한 문장으로 표현하지 못하고 한두 개의 키워드만을 제시하는 게 고작이다. 이러한 상황에서 정보 검색 시스템 혹은 질의응답 시스템이 한번에 사용자가 원하는 답을 찾아주기는 더욱 어렵다. 이런 점을 고려할 때, 정보검색 및 질의응답 시스템 자체의 성능을 향상시키는 것 못지 않게 사용자가 자신의 욕구를 차츰 적절히 수정해 갈 수 있는 대화 형식의 처리 기술이 함께 요구된다.

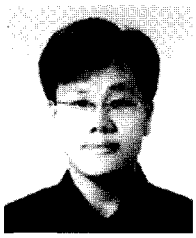
참고문헌

[1] J. Fukumoto, T. Kato, and F. Masui. Question Answering Challenge (QAC-1) Question answering evaluation at NTCIR Workshop 3, Overview Working Notes of the Third NTCIR Workshop Meeting, pp 77-86, Tokyo, 2002. NII.

[2] S. Harabagiu, D. Moldovan, M. Pasca, R. Mihalcea, M. Surdeanu, R. Bunescu, R. Girju, V. Rus, and P. Morarescu. Falcon: Boosting knowledge for answer engines. In

The 9th Text Retrieval Conference (TREC-9), pages 479 488, Maryland, 2000. NIST.
[3] E. Hovy, L. Gerber, U. Hermjakob, M. Junk, and C.-Y. Lin. Question answering in webclopedia. In The 9th Text Retrieval Conference (TREC-9), pages 655-664, Maryland, 2000. NIST.
[4] A. Ittycheriah, M. Franz, W.-J. Zhu, and A. Ratnaparkhi, IBM's Statistical Question Answering System, In The 9th Text Retrieval Conference (TREC-9), pages 229-234, Maryland, 2000. NIST.
[5] M. Kaszkiel and J. Zobel. Passage retrieval revisited. In The 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 178-185, Philadelphia, 1997. ACM.
[6] H. Kim, K. Kim, G. G. Lee, and J. Seo. MAYA: A Fast Question-answering System Based On A Predictive Answer Indexer, in Proceedings of the ACL Workshop Open-Domain Question Answering, pp. 9-16, 2001.
[7] G. G. Lee, J. Seo, S. Lee, H. Jung, B.-H. Cho, C. Lee, B.-K. Kwak, J. Cha. D. Kim, J. An, H. Kim, and K. Kim. SiteQ: Engineering high performance QA system using lexico-semantic pattern matching and shallow NLP. In The 10th Text Retrieval Conference (TREC-10), pages 437-446, Maryland, 2001. NIST.
[8] S. Lee and G. G. Lee. SiteQ/J: A Question Answering System for Japanese. QAC1 Working Notes of the Third NTCIR Workshop Meeting. Pages 31-38, Tokyo, 2002. NII.
[9] S. Lee and G. G. Lee. Use of Dynamic Passage Selection and Lexico-Semantic Patterns for Japanese Natural Language Question Answering. IEICE Transactions on Information and Systems, Vol. E86-D, No.9, pages 1638-1647, 2003.9.
[10] G. Miller. WordNet: A Lexical Database for English, Communications of the ACM 38(11) pp. 39-41, 1995.
[11] J. Prager. Question-answering by predictive annotation. In The 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 184-191, Athens, 2000. ACM.
[12] S. E. Roberston et al. Okapi at TREC-3. In Overview of the Third Text Retrieval Conference (TREC-3), pages 109-126, Maryland, 1995.
[13] M. M. Soubbotin. Patterns of potential answer Expressions as clues to the right answers. In The 10th Text Retrieval Conference (TREC-10), pages 293-302, Maryland, 2001. NIST.
[14] E. M. Voorhees and D. Tice. Building a question answering test collection. In The 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 200-207, Athens, 2000. ACM.
[15] J. Zobel, A. Moffat, R. Wilkinson, and R. Sacks Davis. Efficient retrieval of partial documents. Information Processing and Management, 31(3):361-377, 1995.
[16] Defense Advanced Research Projects Agency. Proceedings of the Seventh Message Understanding-Conference-(MUC-7), 1998. Available at http://www.saic.com.

이 승 우



1997 경북대학교 컴퓨터공학과 학사
1999 포항공과대학교 컴퓨터공학과 석사
1999~2000 포항공과대학교 정보통신연구소 연구원
2001~현재 포항공과대학교 컴퓨터공학과 박사과정
관심분야 : 자연언어처리, 정보검색, 질의응답, 정보추출
E-mail : pinesnow@postech.ac.kr

이 근 배

1984 서울대학교 컴퓨터공학과 학사
1986 서울대학교 컴퓨터공학과 석사
1991 UCLA 컴퓨터학과 박사
1991. 3~1991. 9 UCLA 연구원
1991~1996 포항공과대학교 조교수
1997~2003 포항공과대학교 부교수
2000~2001 미국 Stanford CSLI 연구원
2004~현재 포항공과대학교 정교수
관심분야 : 자연언어 처리, 음성인식, 정보검색, 바이오 인포메틱스
E-mail : gblee@postech.ac.kr